

Problème n° 2

Notations.

$\mathbb{N}$ désigne l'ensemble des entiers naturels et $\mathbb{R}$ l'ensemble des nombres réels.

Pour m et n deux entiers naturels, $\llbracket m, n \rrbracket$ désigne l'ensemble des entiers k tels que $m \leq k \leq n$.

Pour n un entier naturel non nul, $\mathcal{M}_n(\mathbb{R})$ désigne l'ensemble des matrices à n lignes et n colonnes, à coefficients réels.

On note I_n la matrice identité d'ordre n .

Le module d'un nombre complexe z est noté $|z|$.

Définitions.

Soit $(X^{(k)})_{k \geq 0}$ une suite de vecteurs de $\mathbb{R}^n$ et $X = (x_1, \dots, x_n)$ un vecteur de $\mathbb{R}^n$. Pour tout entier naturel k , on pose $X^{(k)} = (x_1^{(k)}, \dots, x_n^{(k)})$. On dit que la suite $(X^{(k)})_{k \geq 0}$ converge vers X si pour tout $i \in \llbracket 1, n \rrbracket$, la suite $(x_i^{(k)})_{k \geq 0}$ converge vers x_i .

Soit $(A^{(k)})_{k \geq 0}$ une suite de matrices de $M_n(\mathbb{R})$ et $A = (a_{i,j})_{1 \leq i,j \leq n}$ une matrice de $M_n(\mathbb{R})$.

Pour tout entier naturel k , on pose $A^{(k)} = (a_{i,j}^{(k)})_{1 \leq i,j \leq n}$. On dit que la suite $(A^{(k)})_{k \geq 0}$ converge vers A si pour tous $i, j \in \llbracket 1, n \rrbracket$, la suite $(a_{i,j}^{(k)})_{k \geq 0}$ converge vers $a_{i,j}$.

Soit G un graphe orienté fini et soient i et j deux sommets de ce graphe. On dit que j est un sommet voisin de i s'il existe une arête orientée de G reliant i à j .

Partie A : marche aléatoire sur un graphe

On considère un graphe orienté fini dont les sommets sont numérotés de 1 à n . Un point se déplace aléatoirement d'un sommet à un autre de ce graphe au cours d'étapes, le nombre d'étapes pouvant tendre vers l'infini. À chaque étape, le point se déplace du sommet où il se trouve vers l'un des sommets voisins de façon équiprobable. Ceci entraîne notamment que la probabilité de passer du sommet i au sommet j ne dépend pas du rang de l'étape.

Pour $1 \leq i, j \leq n$, on note $a_{i,j}$, la probabilité que le point passe du sommet i au sommet j ; en particulier, s'il n'y a pas d'arête reliant i à j , $a_{i,j} = 0$. La matrice dont le coefficient de la ligne i et de la colonne j est égal à $a_{i,j}$ est notée A . Cette matrice s'appelle la *matrice de transition* du graphe.

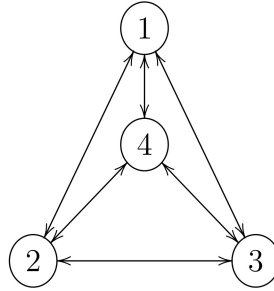
Pour $k \in \mathbb{N}$, on note $P^{(k)}$ le vecteur ligne $(p_1^{(k)}, p_2^{(k)}, \dots, p_{n-1}^{(k)}, p_n^{(k)})$, où pour $1 \leq i \leq n$, $p_i^{(k)}$ est la probabilité que le point soit sur le sommet i à l'étape de rang k .

I. Résultats généraux

1. Justifier que, pour tout entier naturel k , $p_1^{(k)} + \dots + p_n^{(k)} = 1$.
2. Montrer que, pour tout entier naturel k , $P^{(k+1)} = P^{(k)} A$.
3. En déduire, pour tout entier naturel k , une expression de $P^{(k)}$ en fonction de A , k et $P^{(0)}$.
4. On suppose que la suite de vecteurs $(P^{(k)})_{k \geq 0}$ converge vers un vecteur $P = (p_1, \dots, p_n)$. Montrer que $PA = P$ et que $p_1 + \dots + p_n = 1$.

II. Marche aléatoire sur un tétraèdre

Dans cette question, on suppose que G est le graphe ci-dessous :



On remarque que, lorsque le point est sur l'un des sommets du graphe, il a la même probabilité de se rendre sur chacun des trois autres sommets du graphe. On suppose qu'au départ, le point est sur le sommet 1, de sorte que :

$$P^{(0)} = \begin{pmatrix} 1 & 0 & 0 & 0 \end{pmatrix}.$$

On pose :

$$U = \begin{pmatrix} 0 & 1 & 1 & 1 \\ 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 1 \\ 1 & 1 & 1 & 0 \end{pmatrix}.$$

1. Exprimer la matrice de transition A en fonction de U .
2. Calculer U^2 et U^3 .
3. Montrer qu'il existe deux suites $(\alpha_k)_{k \geq 0}$ et $(\beta_k)_{k \geq 0}$ telles que pour tout entier naturel k :

$$U^k = \begin{pmatrix} \alpha_k & \beta_k & \beta_k & \beta_k \\ \beta_k & \alpha_k & \beta_k & \beta_k \\ \beta_k & \beta_k & \alpha_k & \beta_k \\ \beta_k & \beta_k & \beta_k & \alpha_k \end{pmatrix}.$$

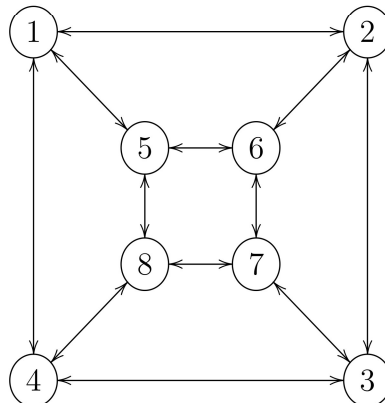
Montrer de plus que, pour tout entier naturel k ,

$$\begin{cases} \alpha_{k+1} = 3\beta_k \\ \beta_{k+1} = \alpha_k + 2\beta_k \end{cases}$$

4. En déduire que, pour tout entier naturel k , $\beta_{k+2} = 2\beta_{k+1} + 3\beta_k$.
5. En déduire que, pour tout entier naturel k , $\beta_k = \frac{3^k - (-1)^k}{4}$ et $\alpha_k = \frac{3^k + 3(-1)^k}{4}$.
6. En déduire pour tout entier naturel k une expression de $P^{(k)}$.
7. Montrer que la suite de vecteurs $(P^{(k)})_{k \geq 0}$ converge et déterminer la limite de $P^{(k)}$ lorsque k tend vers $+\infty$.

III. Marche aléatoire sur une pyramide tronquée à base carrée

Dans cette question, on suppose que G est le graphe ci-dessous :



On rappelle que, lorsque le point est sur l'un des sommets du graphe, il a la même probabilité de se rendre sur chacun des sommets à qui il est relié. On suppose qu'au départ, le point est sur le sommet 1, de sorte que :

$$P^{(0)} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}.$$

On note $X = \{1, 3, 6, 8\}$ et $Y = \{2, 4, 5, 7\}$.

1. Donner la matrice de transition A de ce graphe et calculer

$$\begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \end{pmatrix} A.$$

2. Montrer que si le point se trouve sur un sommet de la partie X à une étape donnée, il se trouvera sur un sommet de Y à l'étape suivante, et que s'il se trouve sur un sommet de Y à une étape donnée, il se trouvera sur un sommet de X à l'étape suivante.
3.
 - a. Démontrer que les coefficients de $P^{(k)}$ dont les indices sont des éléments de X sont nuls si k est impair, et que les coefficients de $P^{(k)}$ dont les indices sont des éléments de Y sont nuls si k est pair.
 - b. La suite de vecteurs $(P^{(k)})_{k \geq 0}$ converge t-elle ?

IV. On revient au cas général d'un graphe G à n sommets. Parmi les trois liens logiques condition nécessaire, condition suffisante et condition nécessaire et suffisante, quel est celui qui relie les deux propositions suivantes ?

- (i) la suite de vecteurs $(P^{(k)})_{k \geq 0}$ converge ;
- (ii) il existe un vecteur $P = (p_1, \dots, p_n)$ de $\mathbb{R}^n$, avec $p_1, \dots, p_n$ positifs ou nuls et $p_1 + \dots + p_n = 1$, tel que $PA = P$.

La réponse, qui devra être soigneusement justifiée, sera présentée sous deux formes : une phrase rédigée en français et une proposition mathématique comportant une implication ou une équivalence.

Partie B : matrices stochastiques et densités de probabilité

Définitions.

Soit $X = (x_1, \dots, x_n)$ un vecteur de $\mathbb{R}^n$. On dit que X est une densité de probabilité si pour tout $i \in \llbracket 1, n \rrbracket$, $x_i \geq 0$ et $x_1 + \dots + x_n = 1$.

Soit A une matrice de $M_n(\mathbb{R})$. On dit que A est une matrice stochastique si chaque ligne de A est une densité de probabilité.

- I. Soit $A \in M_n(\mathbb{R})$, dont tous les coefficients sont positifs ou nuls. Montrer que A est une matrice stochastique si et seulement si

$$A \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}.$$

- II. Montrer que la matrice de transition d'un graphe (définie dans la partie A.) est une matrice stochastique et que, pour tout entier naturel k , le vecteur $P^{(k)}$, lui aussi défini dans la partie A., est une densité de probabilité.

- III. Soit $A \in M_n(\mathbb{R})$ une matrice stochastique et $X \in \mathbb{R}^n$ une densité de probabilité. Montrer que XA est une densité de probabilité.
- IV. Soient A et B deux matrices stochastiques de $M_n(\mathbb{R})$.
1. Montrer que AB est une matrice stochastique.
 2. Soit $\alpha \in [0, 1]$. Montrer que $\alpha A + (1 - \alpha)B$ est une matrice stochastique.
- V. Soit $(X^{(k)})_{k \geq 0}$ une suite de vecteurs de $\mathbb{R}^n$, convergeant vers un vecteur X . Montrer que, si pour tout entier naturel k , $X^{(k)}$ est une densité de probabilité, alors X est une densité de probabilité.
- VI. Soit $(A^{(k)})_{k \geq 0}$ une suite de matrices de $M_n(\mathbb{R})$, convergeant vers une matrice A . Montrer que, si pour tout entier naturel k , $A^{(k)}$ est une matrice stochastique, alors A est une matrice stochastique.

Partie C : PageRank, un premier modèle

Le moteur de recherche Google annonce environ 339 000 résultats à la requête « capes mathématiques ». Sur la première page de résultats, il classe en premier le site `capex-math.org`, puis le site `www.devenirenseignant.gouv.fr`. Le succès de ce moteur de recherche tient à sa rapidité, son exhaustivité et la qualité du classement qu'il effectue entre les différentes pages, à partir d'une « mesure d'importance » attribuée à chacune. Cette mesure d'importance d'une page s'appelle sa pertinence. On se propose d'étudier deux algorithmes permettant de déterminer la pertinence de chaque page du web. Le second algorithme, connu sous le nom de *PageRank*, a été inventé par Sergey Brin et Larry Page.

Le mot « PageRank » lui-même est une marque déposée par Google. Le procédé a été breveté par l'université de Stanford qui en a accordé une licence exclusive à Google.

On modélise le web par un graphe orienté à n sommets représentant chacun une page et dont les arêtes représentent les liens entre celles-ci (liens hypertextes, citations d'articles, etc.) ; la notation $i \rightarrow j$ traduit le fait que la page i pointe vers la page j . Pour tout entier $i \in \llbracket 1, n \rrbracket$, λ_i désigne le nombre de liens issus de la page i et μ_i désigne la pertinence de cette page.

On impose à la suite $(\mu_j)_{1 \leq j \leq n}$ des pertinences des pages du web de vérifier les deux conditions suivantes :

- (i) la pertinence d'une page dépend des pertinences des pages qui pointent vers elle de manière affine ;
- (ii) plus il y a de liens issus d'une page, plus la contribution de cette page dans la pertinence des pages vers lesquelles elle pointe est faible.

- I. Montrer que toute suite $(\mu_j)_{1 \leq j \leq n}$ de nombres réels positifs ou nuls satisfaisant aux n conditions

$$(*) \quad \forall j \in \llbracket 1, n \rrbracket, \quad \mu_j = \sum_{i \rightarrow j} \frac{1}{\lambda_i} \mu_i.$$

vérifie les conditions (i) et (ii).

Toute solution $(\mu_j)_{1 \leq j \leq n}$ du système linéaire $(*)$ formée de réels positifs ou nuls fournit les pertinences des n pages du web. Dans la pratique, la résolution de ce système est impossible, le nombre n de pages du web étant beaucoup trop élevé. Pour cette raison, on cherche une autre méthode pour obtenir les pertinences.

- II.** Soient $i, j \in \llbracket 1, n \rrbracket$.
 – Si $i \neq j$, on pose

$$a_{i,j} = \begin{cases} 0 & \text{s'il n'y a pas de lien de la page } i \text{ vers la page } j ; \\ \frac{1}{\lambda_i} & \text{si } i \rightarrow j. \end{cases}$$

- Si $i = j$, dans le cas où il n'y a aucun lien qui permette de sortir de la page i , on convient d'écrire $a_{i,i} = 1$, pour indiquer que si on arrive sur cette page, on y reste ; sinon, on note $a_{i,i} = 0$.

Pour $1 \leq i, j \leq n$, le coefficient $a_{i,j}$ peut s'interpréter comme la probabilité qu'une personne navigant sur le web et se trouvant sur la page i suive le lien vers la page j en choisissant au hasard entre les λ_i liens disponibles à partir de la page i s'il y en a et reste sur la page i sinon. Soit A la matrice de $M_n(\mathbb{R})$ dont le coefficient de la ligne i et de la colonne j est $a_{i,j}$.

Montrer que A est une matrice stochastique.

- III.** Montrer que la suite finie $(\mu_j)_{1 \leq j \leq n}$ vérifie les conditions (*) introduites à la question **I.** si et seulement si la matrice ligne $M = (\mu_1, \mu_2, \dots, \mu_n)$ vérifie $M = MA$.
- IV.** On note dans la suite $P^{(k)}$ la densité de probabilité après k clics : le coefficient d'indice i de la matrice ligne $P^{(k)}$ est la probabilité que le « surfeur » se trouve sur la page i après k clics.
1. Vérifier que pour tout entier naturel k , $P^{(k+1)} = P^{(k)}A$.
 2. Montrer que si la suite $(P^{(k)})_{k \geq 0}$ converge vers une densité de probabilité limite P , alors cette densité de probabilité vérifie les conditions (*) et fournit donc une mesure de la pertinence des pages du web.

Partie D : PageRank, un second modèle

On a vu dans la partie **A.** que l'existence de cette densité de probabilité limite n'est pas toujours assurée. De plus, il peut arriver que certaines pages ne comportent aucun lien vers d'autres pages. Selon le premier modèle, lorsque le surfeur arrive sur l'une d'entre elles, il lui est impossible de la quitter. Ce modèle n'étant pas conforme à la réalité, on étudie un second modèle qui introduit la possibilité de quitter à chaque clic une page quelconque pour se diriger vers une autre choisie au hasard, et ce avec une probabilité fixe $\alpha \in]0, 1[$, appelée *facteur d'amortissement*.

Non seulement ce modèle permet de mieux rendre compte de la réalité, mais en plus on va montrer qu'il garantit l'existence d'une densité de probabilité limite fournissant une mesure de la pertinence des n pages.

On considère la matrice

$$B = (1 - \alpha)A + \frac{\alpha}{n}J$$

où J est la matrice de $M_n(\mathbb{R})$ dont tous les coefficients sont égaux à 1, A la matrice stochastique définie dans la partie **C.** et L la matrice ligne possédant n coefficients tous égaux à 1 :

$$J = \begin{pmatrix} 1 & \dots & 1 \\ \vdots & & \vdots \\ 1 & \dots & 1 \end{pmatrix}, \quad L = (1, \dots, 1).$$

- I.** Montrer que B est une matrice stochastique.
- II.** On suppose qu'il existe un vecteur ligne $N = (\nu_1, \dots, \nu_n)$ de réels positifs ou nuls vérifiant $N = NB$, tel que $\nu_1 + \dots + \nu_n = 1$. Montrer que la suite $(\nu_1, \dots, \nu_n)$ vérifie les conditions (i) et (ii) énoncées dans la partie **C**.
- III.** Montrer que, si Q est une densité de probabilité, alors $QJ = L$.
- IV.** Justifier que la matrice colonne U de taille n dont tous les coefficients sont égaux à 1 vérifie $AU = U$. En déduire que 1 est une valeur propre de la matrice A .
- V.** Soit $\lambda \in \mathbb{C}$ une valeur propre de A et $Z = \begin{pmatrix} z_1 \\ \vdots \\ z_n \end{pmatrix}$, où $z_i \in \mathbb{C}$ pour tout $i \in \llbracket 1, n \rrbracket$, un vecteur propre associé. On note

$$|Z| = \max_{1 \leq j \leq n} |z_j|.$$

1. À l'aide de l'inégalité triangulaire appliquée à une coordonnée bien choisie du vecteur AZ , montrer que $|\lambda| \leq 1$.
 2. En déduire que, si β est un nombre réel strictement supérieur à 1, la matrice $A - \beta I_n$ est inversible.
 3. En déduire que si γ est un nombre réel strictement compris entre 0 et 1, la matrice $I_n - \gamma A$ est inversible.
- VI.** 1. Justifier que la matrice $I_n - (1 - \alpha)A$ est inversible. Cela permet de poser

$$H = \frac{\alpha}{n} L(I_n - (1 - \alpha)A)^{-1}.$$

2. Vérifier que $H = (1 - \alpha)HA + \frac{\alpha}{n}L$.
- VII.** Soit Q une densité de probabilité. On définit la suite $(Q^{(k)})_{k \geq 0}$ par $Q^{(k)} = QB^k$ pour tout entier naturel k .
1. Démontrer, pour tout entier naturel k , l'égalité $(Q^{(k+1)} - H) = (1 - \alpha)(Q^{(k)} - H)A$.
 2. En déduire que, pour tout entier naturel k ,

$$Q^{(k)} - H = (1 - \alpha)^k (Q - H)A^k.$$

- VIII.** En utilisant le fait que A^k est une matrice stochastique, montrer que les suites formées par les coefficients des matrices lignes $(Q - H)A^k$ sont bornées. Que peut-on en déduire pour la suite des densités de probabilité $Q^{(k)}$?
- IX.** Que peut-on en conclure pour le vecteur ligne H ?
- X.** Quel est l'avantage de cette méthode par rapport à celle exposée dans la partie **C** ?
- XI.** Le calcul de l'inverse de la matrice $I_n - (1 - \alpha)A$ permettant de calculer H étant très coûteux, on se contente d'approcher H par $Q^{(k)} = QB^k$ pour des valeurs de k suffisamment grandes. Quelle justification mathématique peut-on apporter à cette pratique ?